

PCT

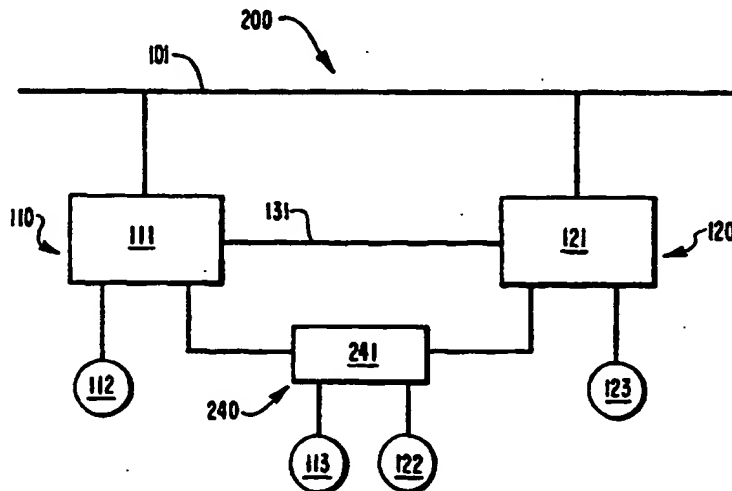
WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau



INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification ⁵ : G06F 11/34		(11) International Publication Number: WO 95/00906
A1		(43) International Publication Date: 5 January 1995 (05.01.95)
(21) International Application Number: PCT/US94/07009		(81) Designated States: AT, AU, BB, BG, BR, BY, CA, CH, CN, CZ, DE, DK, ES, FI, GB, HU, JP, KP, KR, KZ, LK, LU, LV, MG, MN, MW, NL, NO, NZ, PL, PT, RO, RU, SD, SE, SK, UA, US, UZ, VN, European patent (AT, BE, CH, DE, DK, ES, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, ML, MR, NE, SN, TD, TG).
(22) International Filing Date: 21 June 1994 (21.06.94)		
(30) Priority Data: 08/081,391 23 June 1993 (23.06.93) US		
(71) Applicant: VINCA CORPORATION [US/US]; 4000 Central Park East, 1815 South State Street, Orem, UT 84058 (US).		
(74) Agents: CHRISTIANSEN, Jon, C. et al.; Van Cott, Bagley, Cornwall & McCarthy, 50 South Main Street, Suite 1600, Salt Lake City, UT 84144 (US).		

(54) Title: **METHOD FOR IMPROVING DISK MIRRORING ERROR RECOVERY IN A COMPUTER SYSTEM INCLUDING AN ALTERNATE COMMUNICATION PATH**



(57) Abstract

A method for reducing the time necessary to recover from a processor (111, 121) failure in a fault-tolerant computer system with redundant server computer systems (110, 120) with their own disk storage systems is disclosed and claimed. In normal operation whenever data is to be written to disk storage, each of the servers writes an identical copy of the data to its own disk storage system. When a server processor fails and then is restored to operation, that server's disk storage system must be made identical to (consistent with) the disk storage system of the non-failing server before the system is again fault tolerant. This method improves performance by electronically transferring the disk storage system from the failing server to a non-failing server, having the non-failing server keep the transferred disk storage system identical to its normal disk storage system, and reconnecting the transferred disk storage system to the failed server when it again becomes available. This minimizes the processing time required to make the disk storage contents identical, both at the time of failure and at the time of restoration.

Best Available Copy

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AT	Austria	GB	United Kingdom	MR	Mauritania
AU	Australia	GE	Georgia	MW	Malawi
BB	Barbados	GN	Guinea	NR	Niger
BE	Belgium	GR	Greece	NL	Netherlands
BF	Burkina Faso	HU	Hungary	NO	Norway
BG	Bulgaria	IE	Ireland	NZ	New Zealand
BJ	Benin	IT	Italy	PL	Poland
BR	Brazil	JP	Japan	PT	Portugal
BY	Belarus	KE	Kenya	RO	Romania
CA	Canada	KG	Kyrgyzstan	RU	Russian Federation
CF	Central African Republic	KP	Democratic People's Republic of Korea	SD	Sudan
CG	Congo	KR	Republic of Korea	SE	Sweden
CH	Switzerland	KZ	Kazakhstan	SI	Slovenia
CI	Côte d'Ivoire	LI	Liechtenstein	SK	Slovakia
CM	Cameroon	LK	Sri Lanka	SN	Senegal
CN	China	LU	Luxembourg	TD	Chad
CS	Czechoslovakia	LV	Latvia	TG	Togo
CZ	Czech Republic	MC	Monaco	TJ	Tajikistan
DE	Germany	MD	Republic of Moldova	TT	Trinidad and Tobago
DK	Denmark	MG	Madagascar	UA	Ukraine
ES	Spain	ML	Mali	US	United States of America
FI	Finland	MN	Mongolia	UZ	Uzbekistan
FR	France			VN	Viet Nam
GA	Gabon				

METHOD FOR IMPROVING DISK MIRRORING ERROR RECOVERY IN A COMPUTER SYSTEM
INCLUDING AN ALTERNATE COMMUNICATION PATH

SPECIFICATION

To all whom it may concern:

Be it known that Richard Rollins, Michael Ohran, Randall C. Johnson, Scott Bonsteel, and Richard S. Ohran, citizens of the United States of America, have invented a new and useful invention entitled METHOD FOR IMPROVING ERROR RECOVERY PERFORMANCE IN A FAULT-TOLERANT COMPUTER SYSTEM of which the following comprises a complete specification.

1 METHOD FOR IMPROVING ERROR RECOVERY PERFORMANCE
2 IN A FAULT-TOLERANT COMPUTER SYSTEM

3
4 Microfiche Appendix. This specification includes a
5 Microfiche Appendix which includes 1 page of
6 microfiche and a total of 13 frames. The
7 Microfiche Appendix includes computer source code
8 illustrative of one preferred embodiment of the
9 present invention.

10
11 Background of the Invention

12 Field of the Invention. This invention relates to
13 fault-tolerant computer systems, and in particular
14 to the methods used to recover from a computer
15 failure in a system with redundant computers each
16 with its own mass storage system(s).

17 Description of Related Art. It is often desirable
18 to provide continuous operation of computer
19 systems, particularly file servers which support a
20 number of user workstations or personal computers
21 on a network. To achieve this continuous

1 operation, it is necessary for the computer system
2 to be tolerant of software and hardware problems or
3 faults. This is generally done by having redundant
4 computers and redundant mass storage systems, such
5 that a backup computer or disk drive is immediately
6 available to take over in the event of a fault.

7 A number of techniques for implementing a
8 fault-tolerant computer system are described in
9 Major et al., United States Patent 5,157,663, which
0 is hereby incorporated by reference in its
1 entirety, and Major's cited references. In
2 particular, the invention of Major provides a
3 replicated network file server capable of
4 recovering from the failure of either the computer
5 or the mass storage system of one of the two file
6 servers. It has been used by Novell to implement
7 its SFT-III fault-tolerant file server product.

8 Figure 1 illustrates the hardware
9 configuration for a fault-tolerant computer system
10 100, such as described in Major. There are two
11 server computer systems 110 and 120 connected to

1 network 101, from which they receive requests from
2 client computers. While we refer to computers 110
3 and 120 as "server computer systems" or simply
4 "servers" and show them in that role in the
5 examples herein, this should not be regarded as
6 limiting the present invention to computers used
7 only as servers for other computer systems.

8 Server computer system 110 has computer
9 111 which includes a central processing unit and
10 appropriate memory systems and other peripherals.
11 Server computer system 120 has computer 121 which
12 includes a central processing unit and appropriate
13 memory systems and other peripherals. Mass storage
14 systems 112 and 113 are connected to computer 111,
15 and mass storage systems 122 and 123 are connected
16 to computer 121. Mass storage systems 112 and 123
17 are optional devices for storing operating system
18 routines and other data not associated with read
19 and write requests received from network 101.
20 Finally, there is an optional communications link
21 131 between computers 111 and 121.

1 The mass storage systems can be
2 implemented using magnetic disk drives, optical
3 discs, magnetic tape drives, or any other medium
4 capable of handling the read and write requests of
5 the particular computer system.

6 An operating system or other control
7 program runs on server computer systems 110 and
8 120, executed by computers 111 and 121,
9 respectively. This operating system handles server
0 requests received from network 101 and controls
1 mass storage systems 112 and 113 on server 110, and
2 mass storage systems 122 and 123 on server 120, as
3 well as any other peripherals attached to computers
4 111 and 121.

5 While Figure 1 illustrates only two
6 server computer systems 110 and 120, because that
7 is the most common (and lowest cost) configuration
8 for a fault-tolerant computer system 100,
9 configurations with more than two server computer
0 systems are possible and do not depart from the
1 spirit and scope of the present invention.

1 In normal operation, both server computer
2 system 110 and server computer system 120 handle
3 each mass storage write request received from
4 network 101. Server computer system 110 writes the
5 data from the network request to mass storage
6 system 113, and server computer system 120 writes
7 the data from the network request to mass storage
8 system 122. This results in the data on mass
9 storage system 122 being the mirror image of the
10 data on mass storage system 113 and the states of
11 server computer systems 110 and 120 are generally
12 consistent. In the following discussion, the
13 process of maintaining two or more identical copies
14 of information on separate mass storage systems is
15 referred to as "mirroring the information".

16 (For read operations, either server
17 computer system 110 or server computer system 120
18 can handle the request without involving the other
19 server, since a read operation does not change the
20 state of the information stored on the mass storage
21 systems.)

1 Although computer system 100 provides a
2 substantial degree of fault tolerance, when one of
3 server computer systems 110 or 120 fails, the fault
4 tolerance of the system is reduced. In the most
5 common case of two server computer systems, as
6 illustrated by Figure 1, the failure of one server
7 computer system results in a system with no further
8 tolerance to hardware faults or many software
9 faults.

0 In a fault-tolerant computer system such
1 as described above, it is necessary after a failed
2 server computer system has been restored to bring
3 the previously-failed computer system into a state
4 consistent with the server computer system that has
5 continued operating. This requires writing all the
6 changes made to the mass storage system of the non-
7 failing server to the mass storage system of the
8 previously-failed server so that the mass storage
9 systems again mirror each other. Until that has
10 been accomplished, the system is not fault tolerant
11 even though the failed server has been restored.

1 If a server has been unavailable due to
2 its failure for a period of time during which there
3 have been only a limited number of changes made to
4 the mass storage system of the non-failing server,
5 it is possible for the non-failing server to
6 remember all the changes made (for example, by
7 keeping them in a list stored in its memory) and
8 forward the changes to the previously-failed server
9 when it has been restored to operation. The
10 previously-failed server can then update its mass
11 storage system with the changes and make it
12 consistent with the non-failing server. This
13 process typically does not cause excessive
14 performance degradation to the non-failing server
15 for any substantial period of time.

16 However, if there have been more changes
17 than can be conveniently remembered by the non-
18 failing server, then the non-failing server must
19 transfer all the information from its mass storage
20 system to the previously-failed server for writing
21 on its mass storage system in order to ensure that

1 the two servers are consistent. This is a very
2 time consuming and resource-intensive operation,
3 especially if the non-failing server must also
4 handle server requests from the network while this
5 transfer is taking place. For very large mass
6 storage systems, as would be found on servers
7 commonly in use today, and with a reasonably high
8 network request load, it might be many hours before
9 the mass storage systems are again consistent and
10 the system is again fault tolerant. Additionally,
11 the resource-intensiveness of the recovery
12 operation can cause very substantial performance
13 degradation of the non-failed server in processing
14 network requests.

15 Summary of the Invention

16 It is an object of the present invention
17 to provide tolerance to disk faults even though the
18 computer of a server computer system has failed.
19 This is achieved by electronically switching the
20 mass storage system used for network requests from
21 the failed server computer system to the non-

1 failing server computer system. After the mass
2 storage system from the failed server computer
3 system has been connected to the non-failing
4 server's computer, it is made consistent with the
5 mass storage system of the non-failing server.
6 This is typically a quick and simple operation.
7 From that point on, the mass storage system from
8 the failed server it is operated as a mirrored disk
9 system, with each change being written by the non-
10 failing server's computer to both the non-failing
11 server's original mass storage system and to the
12 mass storage system previously on the failed server
13 and now connected to the non-failing server's
14 computer.

15 While operating in this mode, the system
16 will no longer be tolerant to processor failures if
17 the non-failing server is the only remaining server
18 (as would be the case in the common two-server
19 configuration described above), but the system
20 would be tolerant to failures of one of the mass
21 storage systems.

1 It is a further object of the present
2 invention to minimize the time the system is not
3 fault tolerant by eliminating the need for time-
4 consuming copying of the information stored on the
5 mass storage system of the non-failing server to
6 the mass storage of the previously-failed server to
7 make the two mass storage systems again consistent
8 and permit mirroring of information again.

9 This is also achieved by electronically
10 switching the mass storage system from the failed
11 server computer system to the non-failing server
12 computer system. If this switch is accomplished
13 after there have been only a small number of
14 changes to the mass storage system of the non-
15 failing server, the mass storage system from the
16 failed server computer system can be quickly
17 updated and made consistent, allowing mirroring to
18 resume.

19 Furthermore, since the mirroring of the
20 invention keeps the information on the mass storage
21 system from the failed server consistent while it

1 is connected to the non-failing sever computer
2 system, when the mass storage system is reconnected
3 to the previously-failed server only those changes
4 made between the time it was disconnected from the
5 non-failed server and when it becomes available on
6 the previously-failed server need to be made before
7 it is again completely consistent and mirroring by
8 the two servers (and full fault tolerance) resumes.
9 This results in avoiding the substantial
0 performance degradation experienced by the non-
1 failing server during recovery using the prior art
2 recovery method described above. As a result, the
3 invention provides rapid recovery from a fault in
4 the system.

5 These and other features of the invention
6 will be more readily understood upon consideration
7 of the attached drawings and of the following
8 detailed description of those drawings and the
9 presently preferred embodiments of the invention.

10 Brief Description of the Drawings

1 Figure 1 illustrates a prior art
2 implementation of a fault-tolerant computer system
3 with two server computer systems.

4 Figure 2 illustrates the fault-tolerant
5 computer system of Figure 1, modified to permit the
6 method of the invention by including means for
7 connecting a mass storage system to either server's
8 computer.

9 Figure 3 is a flow diagram illustrating
10 the steps to be taken when a processor failure is
11 detected.

12 Figure 4 is a flow diagram illustrating
13 the steps to be taken when the previously-failed
14 processor becomes available.

15 Detailed Description of the Invention

16 Referring to fault-tolerant computer
17 system 200 of Figure 2, and comparing it to prior
18 art fault-tolerant computer system 100 as
19 illustrated in Figure 1, we see that mass storage
20 systems 113 and 122, which were used for storing
21 the information read or written in response to

1 requests from other computer systems on network
2 101, are now part of reconfigurable mass storage
3 system 240. In particular, mass storage system 113
4 can be selectively connected by connection means
5 241 to either computer 111 or computer 121 (or
6 possibly both computers 111 and 121, although such
7 dual connection is not necessary for the present
8 invention), and mass storage system 122 can
9 likewise be independently selectively connected to
10 either computer 111 or computer 121 by connection
11 means 241. The mass storage system 240 is
12 reconfigurable because of the ability to select and
13 change connections between mass storage devices and
14 computers.

15 While Figure 2 illustrates the most
16 common dual server configuration anticipated by the
17 inventors, other configurations with more than two
18 servers are within the scope of the present
19 invention, and the extension of the techniques
20 described below to other configurations will be
21 obvious to one skilled in the art.

1 There are a number of ways such
2 connection means 241 can be implemented, depending
3 on the nature of the mass storage system interface
4 to computers 111 or 121. Connection means 241 can
5 be two independent two-channel switches, which
6 electronically connect all the interface signals
7 from a mass storage system to two computers. Such
8 two-channel switches may be a part of the mass
9 storage system (as is common for mass storage
10 systems intended for use with mainframe computers)
11 or can be a separate unit. A disadvantage of using
12 two-channel switches is the large number of
13 switching gates that are necessary if the number of
14 data and control lines in the mass storage
15 interface is large. That number increases rapidly
16 when there are more than two server computer
17 systems in fault-tolerant computer system 200. For
18 example, a fault-tolerant computer system with
19 three computers connected to three mass storage
20 systems would require 2.25 times the number of
21 switching gates as the system illustrated in Figure

1 2. (The number of switching gates is proportional
2 to the number of computers times the number of mass
3 storage systems.) The number of switching gates
4 can be reduced by not connecting every mass storage
5 system to every computer, although such a
6 configuration would be less flexible in its
7 reconfiguration ability.

8 Another implementation of connection
9 means 241 is for both computer 111 and computer 121
10 to have interfaces to a common bus to which mass
11 storage systems 113 and 122 are also connected. An
12 example of such a bus is the small computer system
13 interface (SCSI) as used on many workstations and
14 personal computers. When a computer wishes to
15 access a mass storage system, the computer requests
16 ownership of the bus through an appropriate bus
17 arbitration procedure, and when ownership is
18 granted, the computer performs the desired mass
19 storage operation. A disadvantage of this
20 implementation is that only one computer (the one

1. with current bus ownership) can access a mass
2 storage system at a time.

3. If it is desirable to use a standard SCSI
4 bus as means 241 for connecting mass storage
5 systems 113 and 122 to computers 111 and 121, and
6 to allow simultaneous access of the mass storage
7 systems 113 and 122 by their respective server's
8 computers, computers 111 and 121 can each have two
9 SCSI interfaces, one connected to mass storage
0 system 113 and one connected to mass storage system
1 122. Mass storage system 113 will be on a SCSI bus
2 connected to both computers 111 and 121, and mass
3 storage system 122 will be on a second SCSI bus,
4 also connected to both computers 111 and 121. If
5 computer 111 or computer 121 is not using a
6 particular mass storage system, it will configure
7 its SCSI interface to be inactive on that mass
8 storage systems particular bus.

9. In the preferred embodiment, a high-speed
0 serial network between computers 111 and 121 and
1 mass storage systems 113 and 122 forms connection

1 means 241. Each computer 111 contains an interface
2 to the network, and requests to a mass storage
3 system 113 or 122 are routed to the appropriate
4 network interface serving the particular mass
5 storage system. Although a bus-type network, such
6 as an Ethernet, could be used, the network of the
7 preferred embodiment has network nodes at each
8 computer and at each mass storage system. Each
9 node can be connected to up to four other network
10 nodes. A message is routed by each network node to
11 a next network node closer to the message's final
12 destination.

13 For the fault-tolerant computer system
14 configuration of Figure 2, one network connection
15 from the node at computer 111 is connected to the
16 node for mass storage system 113, and another
17 network connection from the node at computer 111 is
18 connected to the node for mass storage system 122.
19 Similar connections are used for computer 121.
20 Mass storage system 113's node is connected
21 directly to the nodes for computers 111 and 121,

1 and mass storage system 122's node is similarly
2 connected (but with different links) to computers
3 111 and 121. Routing of messages is trivial, since
4 there is only one link between each computer and
5 each mass storage system.

6 The particular connecting means 241 used
7 to connect computers 111 and 121 to mass storage
8 systems 113 and 122 is not critical to the method
9 of the present invention, so long as it provides
10 for the rapid switching of a mass storage system
11 from one computer to another without affecting the
12 operation of the computers. Any such means for
13 connecting a mass storage system to two or more
14 computers is usable by the method of the present
15 invention.

16 The method of the present invention is
17 divided into two portions, a first portion for
18 reacting to a processor failure and a second
19 portion for recovering from a processor failure.
20 The first portion of the method of the present
21 invention is illustrated by Figure 3, which is a

1 flow diagram illustrating the steps to be taken
2 when a processor failure is detected. The
3 description of the method provided below should be
4 read in light of Figure 2. For purposes of
5 illustration, it will be assumed that connection
6 means 241 initially connects mass storage system
7 113 to computer 111 and mass storage system 122 to
8 computer 121, providing an equivalent to the
9 configuration illustrated in Figure 1 although the
10 connection means 241 of Figure 2 facilitates this
11 equivalent configuration. Information mirroring as
12 described above is being performed by computers 111
13 and 122. It is also assumed that computer 121 has
14 experienced a fault, causing server computer system
15 120 to fail.

16 The method starts in step 301, with each
17 computer 111 and 122 waiting to detect a failure of
18 another server's computer 111 and 122. Such
19 failure can be detected by probing the status of
20 the other server's computer by a means appropriate
21 to the particular operating system being used and

1 the communications methods between the servers. In
2 the case of Novell's SFT-III, the method will be
3 running as a NetWare Loadable Module, or NLM, and
4 be capable of communicating directly with the
5 operating system by means of requests. The NLM
6 will make a null request to the SFT-III process.
7 This null request will be such that it will never
8 normally run to completion, but will remain in the
9 SFT-III process queue. (It will require minimal
10 resources while it remains in the process queue.)
11 In the event of a failure of server computer system
12 121, SFT-III running on server computer system 111
13 will indicate the failure of the null request to
14 the NLM of the method, indicating the failure of
15 server 121. Because a processor failure has been
16 detected, the method depicted in Figure 3 proceeds
17 to step 302.

18 In step 302, detection of the failure of
19 server 121 causes the discontinuation of mirroring
20 information on the failed server 121. This
21 discontinuation can either be done automatically by

1 the operating system upon its detection of the
2 failure of server 121, or by the particular
3 implementation of the preferred embodiment of the
4 method of the present invention. In the case of
5 SFT-III, the discontinuation of mirroring on server
6 121 is performed by the SFT-III operating system.
7 Step 303 of the method is performed next.

8 In step 303, SFT-III remembers all data
9 not mirrored on server 121 following its failure as
0 long as the amount of data to be remembered does
1 not exceed the capacity of the system resource
2 remembering the data. If the particular operating
3 system does not remember non-mirrored data, step
4 303 would have to be performed by the particular
5 implementation of the method of the present
6 invention. The step of remembering all non-
7 mirrored data could be performed by any technique
8 known to persons skilled in the art.

9 Next, step 304 of the method sets
0 connection means 241 to disconnect mass storage
1 system 122 from computer 121 of failed server 120,

1 and to connect it to computer 111 of non-failing
2 server 110. At this point, the method can quickly
3 test mass storage system 122 to determine if it is
4 the cause of the failure of server 120. If it is,
5 there is no fault-tolerance recovery possible using
6 the method, and mass storage system 122 can be
7 disconnected from computer 111 at connection means
8 241. If mass storage system 122 is not the cause
9 of server 120's failure, then the cause must be
10 computer 121, and the method can continue to
11 achieve limited fault tolerance in the presence of
12 the computer 121's failure.

13 Step 305 commands the operating system of
14 server 110 to scan for new mass storage systems,
15 causing the operating system to determine that mass
16 storage system 122 is now connected to computer
17 111, along with mass storage system 113. SFT-III
18 will detect through information on mass storage
19 systems 113 and 122 that they contain similar
20 information, but that mass storage system 122 is
21 not consistent with mass storage system 113. In

1 step 306, SFT-III will update mass storage system
2 122 using the information remembered at step 303
3 and, after the two mass storage systems are
4 consistent (i.e., contain identical mirrored copies
5 of the stored information), step 307 will begin
6 mirroring all information on both mass storage
7 systems 113 and 122 and resume normal operation of
8 the system. If an operating system different than
9 SFT-III does not provide this automatic update for
10 consistency and mirroring, the implementation of
11 the method will have to provide an equivalent
12 service.

13 Note that when SFT-III is used, the only
14 steps of the method that must be performed by the
15 NETWARE loadable module are: (1) detecting the
16 failure of server 120 (step 301), (2) setting
17 communications means 241 to disconnect mass storage
18 system 122 from computer 121 and connecting it to
19 computer 111 (step 304), (3) determining if mass
20 storage system 122 was the cause of the failure of
21 server 120 (also part of step (304), and (4)

1 commanding SFT-III to scan for mass storage systems
2 so that it finds the newly-connected mass storage
3 system 122 (step 305). All the other steps are
4 performed as part of the standard facilities of
5 SFT-III. In other embodiments of the invention,
6 responsibility for performing the steps of the
7 method may be allocated differently.

8 Figure 4 is a flow diagram illustrating
9 the second portion of the invention - the steps to
10 be taken when previously-failed server 120 becomes
11 available again. Server 120 would typically become
12 available after correction of the problem that
13 caused its failure described above. Step 401
14 determines that server 102 is available and the
15 method proceeds to step 402. In step 402, the
16 method sets connection means 241 to disconnect mass
17 storage system 122 from computer 111 after
18 commanding SFT-III on server 110 to remove mass
19 storage system 122 from its active mass storage
20 systems. Due to the unavailability of mass storage
21 system 122 on server 110, data mirroring on server

1 110 will be stopped by SFT-III and it will begin
2 remembering changes to mass storage system 113 not
3 made to mass storage system 122 to be used in
4 making the storage systems consistent later.

5 In step 403, mass storage system 122 is
6 reconnected to computer 121, and in step 404, SFT-
7 III on server 120 is commanded to scan for the
8 newly-connected mass storage system 122. This
9 returns mass storage system 122 to the computer 121
0 to which it was originally connected prior to a
1 server failure. When SFT-III on server 120 detects
2 mass storage system 122, it communicates with
3 server 110 over link 131. At this point, the
4 operating systems on servers 110 and 120 work
5 together to make mass storage system 122 again
6 consistent with mass storage system 113 (i.e., by
7 remembering interim changes to mass storage system
8 113 and writing them to mass storage system 122),
9 and when consistency is achieved, data mirroring on
0 the two servers resumes. At this point, recovery
1 from the server failure is complete.

1 In an SFT-III system, the only steps of
2 the method that the NetWare Loadable Module must
3 perform are: (1) detecting the availability of
4 server 120 (step 401), (2) removing mass storage
5 system 122 from the operating system on server 110
6 (step 402), (3) disconnecting mass storage system
7 122 from computer 111 and connecting it to computer
8 121 by setting connection means 241 (step 403), and
9 (4) commanding SFT-III on server 120 to scan for
10 mass storage so that it locates mass storage system
11 122 (step 404). The steps involved with making
12 mass storage systems 113 and 122 consistent and
13 reestablishing data mirroring (step 405) are
14 performed as part of the standard facilities of
15 SFT-III. In other embodiments of the invention,
16 responsibility for the steps of the method may be
17 allocated differently.

18 Figure 2 illustrates optional mass
19 storage systems 112 and 123 attached to computers
20 111 and 121, respectively. While these two mass
21 storage systems are not required by the method of

1 the present invention, they are useful during the
2 restoration of a failed server. They provide
3 storage for the operating system and other
4 information needed by failed server 120 to begin
5 operation before mass storage system 122 is
6 switched from computer 111 to computer 121. Were
7 mass storage system 123 not available, some means
8 of having mass storage system 122 connected both to
9 computer 121 (for initializing its operation
10 following correction of its failure) and computer
11 111 (for continued disk mirroring) would be
12 necessary. Alternatively, if the initialization
13 time of server 120 is short, mass storage system
14 122 could be switched from computer 111 to computer
15 121 at the start of server 120's initialization,
16 though this would result in more changes that must
17 be remembered and made before data mirroring can
18 begin again.

19 It is to be understood that the above
20 described embodiments are merely illustrative of
21 numerous and varied other embodiments which may

1 constitute applications of the principles of the
2 invention. Such other embodiments may be readily
3 devised by those skilled in the art without
4 departing from the spirit or scope of this
5 invention and it is our intent they be deemed
6 within the scope of our invention.
7

Claims

We claim:

1. A method for rapid failure recovery and system restoration in a fault-tolerant computer system, said computer system comprising:

(A) a first server computer system,
comprising a first computer executing an
operating system;

(B) a second server computer system,
comprising a second computer executing an
operating system;

(C) a first mass storage system connected to said first computer;

(D) a second mass storage system; and

(E) means for connecting said second mass storage system to said first computer and to said second computer;

WHEREIN whenever said first computer writes
data to said first mass storage system, said second

1 computer writes a mirror copy of said data to said
2 second mass storage system,
3 the method comprising the steps of:

4 (1) detecting a failure of said second
5 computer;

6 (2) discontinuing causing said writing of
7 said mirror copy on said second mass storage
8 system;

9 (3) remembering data written to said first
10 mass storage system but not written to said
11 second mass storage system;

12 (4) configuring said second mass storage
13 system to record information from said first
14 computer;

15 (5) writing said remembered data to said
16 second mass storage system;

17 (6) whenever new data is written to said
18 first mass storage system, writing a mirror
19 copy of said new data to said second mass
20 storage system;

1 (7) detecting said second computer's
2 availability;
3 (8) reconfiguring said second mass storage
4 system to record information from said second
5 computer;
6 (9) reestablishing data mirroring such that
7 whenever said first computer writes data to
8 said first mass storage system, said second
9 computer writes a mirror copy of said data on
0 said second mass storage system.

1 2. A method as in claim 1 wherein step (1) is
2 performed by said first computer.

3 3. A method as in claim 2 wherein step (2) is
4 performed by said first computer.

5 4. A method as in claim 1 wherein step (3) is
6 performed by said first computer.

7 5. A method as in claim 4 wherein step (5) is
8 performed by said first computer.

9 6. A method as in claim 5 wherein step (6) is
0 performed by said first computer.

1. 7. A method as in claim 1, wherein said first
2 mass storage system and said second mass storage
3 system each comprise at least one magnetic disk
4 drive.

5 8. A method as in claim 1, wherein said means
6 for connecting said second mass storage system
7 comprises a serial network.

8 9. A method as in claim 1 wherein said operating
9 systems are the SFT-III operating system.

0 10. A method as in claim 9 wherein steps (1), (4)
1 and (5) are performed by a NETWARE loadable module.
2

3 11. A method for rapid failure recovery and
4 system restoration in a fault-tolerant computer
5 system, said computer system comprising:

6 (A) a first server computer system,
7 comprising a first computer executing an
8 operating system;

9 (B) a second server computer system,
0 comprising a second computer executing an
1 operating system;

1 (C) a first mass storage system connected to
2 said first computer;

3 (D) a second mass storage system; and

4 (E) means for selectively connecting said
5 second mass storage system to said first
6 computer and to said second computer;

7 WHEREIN in the absence of a fault said second
8 mass storage system is connected to said second
9 computer; and

0 WHEREIN whenever said first computer writes
1 data to said first mass storage system said first
2 computer can also cause said second computer to
3 write a mirror copy of said data to said second
4 mass storage system,

5 the method of the invention comprising:

6 (1) on said first computer, detecting a
7 failure of said second computer;

8 (2) on said first computer, discontinuing
9 causing said writing of said mirror copy on
0 said second mass storage system by said
1 second computer;

1 (3) on said first computer, remembering data
2 written to said first mass storage system but
3 not written to said second mass storage
4 system;

5 (4) on said first computer, setting said
6 means for connecting said second mass storage
7 system to connect said second mass storage
8 system to said first computer;

9 (5) on said first computer, commanding said
10 operating system of said first computer to
11 scan for mass storage systems such that said
12 operating system of said first computer will
13 determine that both said first mass storage
14 system and said second mass storage system
15 are now connected to said first computer;

16 (6) on said first computer, writing said
17 remembered data to said second mass storage
18 system;

19 (7) on said first computer, whenever new data
20 is written to said first mass storage system,

1 writing a mirror copy of said new data to
2 said second mass storage system;
3 (8) on said first computer, detecting said
4 second computer's availability;
5 (9) on said first computer, commanding said
6 operating system of said first computer to
7 remove said second mass storage system;
8 (10) setting said means for connecting said
9 second mass storage system to connect said
10 second mass storage system to said second
11 computer;
12 (11) on said second computer, commanding
13 said operating system of said second computer
14 to scan for mass storage systems such that
15 said operating system of said second computer
16 will determine that said second mass storage
17 system is now connected to said second
18 computer;
19 (12) reestablishing data mirroring such that
20 whenever said first computer writes data to
21 said first mass storage system said first

1 computer also causes said second computer to
2 write a mirror copy of said data on said
3 second mass storage system.

4 12. A method as in claim 11, wherein said first
5 mass storage system and said second mass storage
6 system each comprise at least one magnetic disk
7 drive.

8 13. A method as in claim 12, wherein said means
9 for connecting said second mass storage system
10 comprises a serial network.

11
12 14. A method for rapid failure recovery in a
13 fault-tolerant computer system, said computer
14 system comprising:

15 (A) a first server computer system,
16 comprising a first computer executing an
17 operating system;

18 (B) a second server computer system,
19 comprising a second computer;

20 (C) a first mass storage system connected to
21 said first computer;

1 (D) a second mass storage system; and

2 (E) means for selectively connecting said
3 second mass storage system to said first
4 computer and to said second computer;

5 WHEREIN in the absence of a fault said second
6 mass storage system is connected to said second
7 computer; and

8 WHEREIN whenever said first computer writes
9 data to said first mass storage system said first
10 computer can also cause said second computer to
11 write a mirror copy of said data on said second
12 mass storage system,

13 the method of the invention comprising said first
14 computer performing the steps of:

15 (1) detecting a failure of said second
16 computer;

17 (2) discontinuing causing said writing of
18 said mirror copy on said second mass storage
19 system by said second computer;

1 (3) remembering data written to said first
2 mass storage system but not written to said
3 second mass storage system;

4 (4) setting said means for connecting said
5 second mass storage system to connect said
6 second mass storage system to said first
7 computer;

8 (5) commanding said operating system of said
9 first computer to scan for mass storage
10 systems such that said operating system of
11 said first computer will determine that both
12 said first mass storage system and said
13 second mass storage system are now connected
14 to said first computer;

15 (6) writing said remembered data to said
16 second mass storage system;

17 (7) whenever new data is written to said
18 first mass storage system, writing a mirror
19 copy of said new data to said second mass
20 storage system.

1 15. A method as in claim 14, wherein said first
2 mass storage system and said second mass storage
3 system each comprise at least one magnetic disk
4 drive.

5 16. A method as in claim 15, wherein said means
6 for connecting said second mass storage system
7 comprises a serial network.

8
9 17. A method for system restoration in a fault-
10 tolerant computer system, said computer system
11 comprising:

12 (A) a first server computer system,
13 comprising a first computer executing an
14 operating system;

15 (B) a second server computer system,
16 comprising a second computer executing an
17 operating system;

18 (C) a first mass storage system connected to
19 said first computer;

20 (D) a second mass storage system; and

1 (E) means for connecting said second mass
2 storage system to said first computer and to
3 said second computer;

4 WHEREIN said second computer is initially
5 unavailable for use, and

6 WHEREIN said second mass storage system is
7 initially connected to said first computer, the
8 method comprising:

9 (1) on said first computer, detecting said
0 second computer's availability;

1 (2) on said first computer, commanding said
2 operating system of said first computer to
3 remove said second mass storage system;

4 (3) setting said means for connecting said
5 second mass storage system to connect said
6 second mass storage system to said second
7 computer;

8 (4) on said second computer, commanding said
9 operating system of said second computer to
10 scan for mass storage systems such that said
11 operating system of said second computer will

1 determine that said second mass storage
2 system is now connected to said second
3 computer;

4 (5) reestablishing data mirroring such that
5 whenever said first computer writes data to
6 said first mass storage system said first
7 computer also causes said second computer to
8 write a mirror copy of said data on said
9 second mass storage system.

10 18. A method as in claim 17, wherein said first
11 mass storage system and said second mass storage
12 system each comprise at least one magnetic disk
13 drive.

14 19. A method as in claim 18, wherein said means
15 for connecting said second mass storage system
16 comprises a serial network.

17 20. A method as in claim 17 wherein said
18 operating system is the SFT-III operating system.

19 21. A method as in claim 20 wherein steps (1),
20 (4) and (5) are performed by a NETWARE loadable
21 module.

1
2 22. A method for rapid failure recovery in a
3 fault-tolerant computer system, said computer
4 system comprising:

5 (A) a first server computer system,
6 comprising a first computer executing an
7 operating system;

8 (B) a second server computer system,
9 comprising a second computer executing an
0 operating system;

1 (C) a first mass storage system connected to
2 said first computer;

3 (D) a second mass storage system; and

4 (E) means for connecting said second mass
5 storage system to said first computer and to
6 said second computer;

7 WHEREIN whenever said first computer writes
8 data to said first mass storage system, said second
9 computer writes a mirror copy of said data to said
10 second mass storage system,
11 the method comprising the steps of:

1 (1) detecting a failure of said second
2 computer;
3 (2) discontinuing causing said writing of
4 said mirror copy on said second mass storage
5 system;
6 (3) remembering data written to said first
7 mass storage system but not written to said
8 second mass storage system;
9 (4) configuring said second mass storage
0 system to record information from said first
1 computer;
2 (5) writing said remembered data to said
3 second mass storage system; and
4 (6) whenever new data is written to said
5 first mass storage system, writing a mirror
6 copy of said new data to said second mass
7 storage system.

8
9 23. A method for system restoration in a fault-
10 tolerant computer system, said computer system
11 comprising:

1 (A) a first server computer system,
2 comprising a first computer executing an
3 operating system;

4 (B) a second server computer system,
5 comprising a second computer executing an
6 operating system;

7 (C) a first mass storage system connected to
8 said first computer;

9 (D) a second mass storage system;

0 (E) means for connecting said second mass
1 storage system to said first computer and to
2 said second computer;

3 WHEREIN said second computer is initially
4 unavailable for use; and

5 WHEREIN said second mass storage system is
6 initially configured to record information from
7 said first computer,

8 the method comprising the steps of:

9 (1) detecting said second computer's
10 availability;

1 (2) reconfiguring said second mass storage
2 system to record information from said second
3 computer;
4 (3) establishing data mirroring such that
5 whenever said first computer writes data to
6 said first mass storage system, said second
7 computer writes a mirror copy of said data on
8 said second mass storage system.
9

0 24. A method for rapid failure recovery and
1 system restoration in a fault-tolerant computer
2 system, the method comprising the steps of:

3 (1) obtaining a computer system, the
4 computer system comprising:

5 (A) a first server computer system,
6 comprising a first computer executing an
7 operating system;

8 (B) a second server computer system,
9 comprising a second computer executing an
0 operating system;

1. (C) a first mass storage system
2 connected to said first computer;
3 (D) a second mass storage system; and
4 (E) means for connecting said second
5 mass storage system to said first
6 computer and to said second computer;
7 (2) operating said computer system such that
8 absent a fault, whenever said first computer writes
9 data to said first mass storage system, said second
0 computer writes a mirror copy of said data to said
1 second mass storage system;
2 (3) detecting a failure of said second
3 computer;
4 (4) discontinuing causing said writing of
5 said mirror copy on said second mass storage
6 system;
7 (5) remembering data written to said first
8 mass storage system but not written to said second
9 mass storage system;

1 (6) configuring said second mass storage
2 system to record information from said first
3 computer;

4 (7) writing said remembered data to said
5 second mass storage system;

6 (8) whenever new data is written to said
7 first mass storage system, writing a mirror copy of
8 said new data to said second mass storage system;

9 (9) detecting said second computer's
10 availability;

11 (10) reconfiguring said second mass storage
12 system to record information from said second
13 computer;

14 (11) reestablishing data mirroring such that
15 whenever said first computer writes data to said
16 first mass storage system, said second computer
17 writes a mirror copy of said data on said second
18 mass storage system.
19

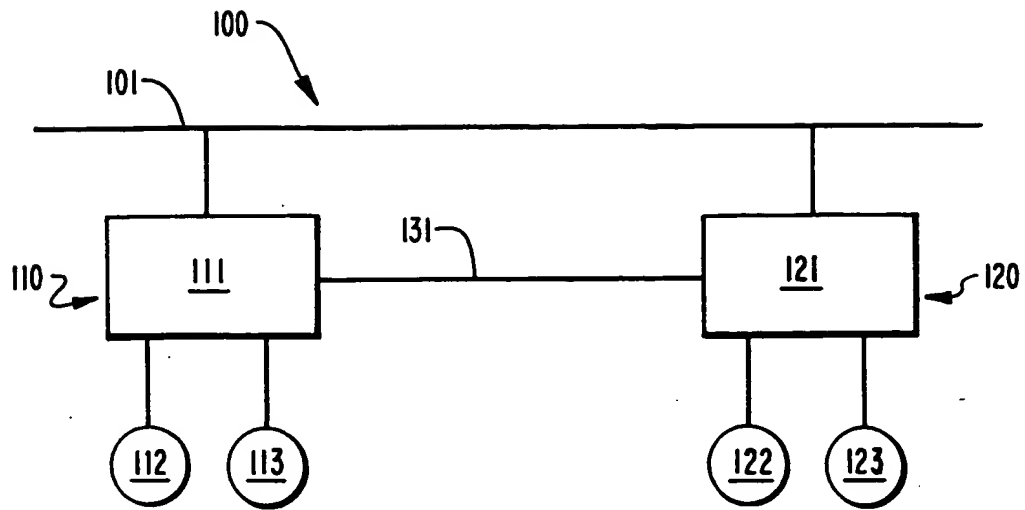


FIG. 1
(PRIOR ART)

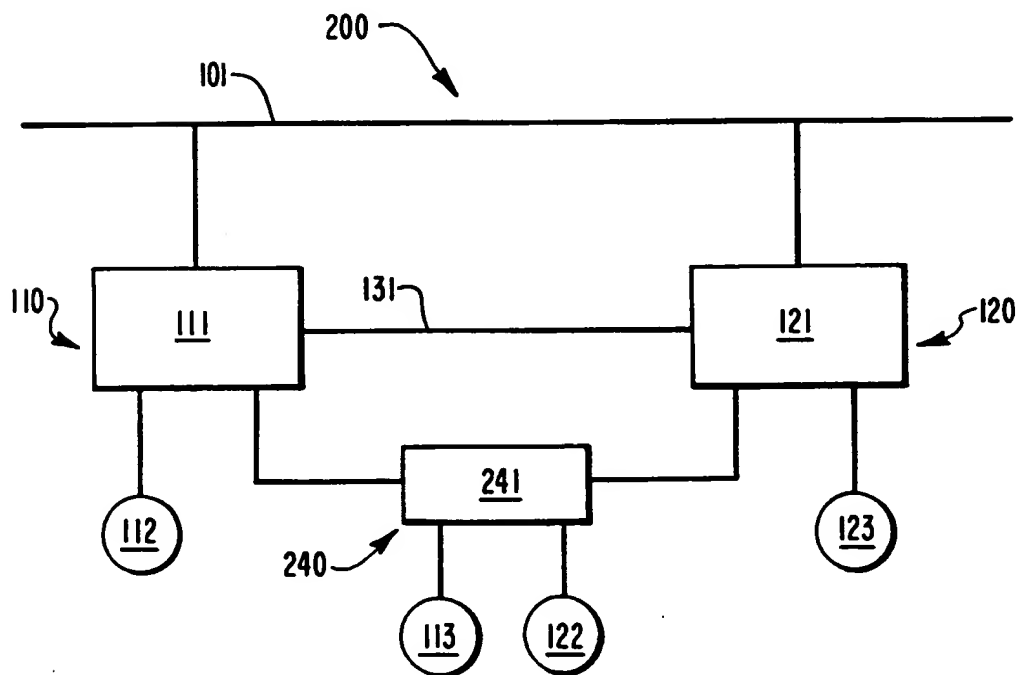


FIG. 2

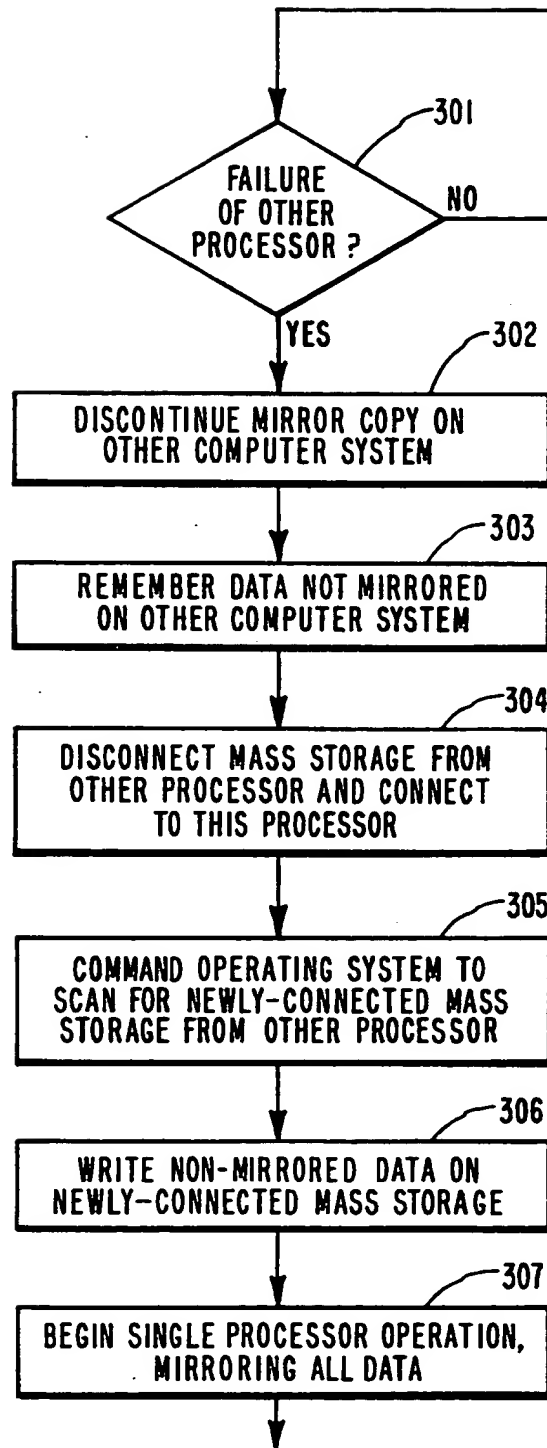


FIG. 3

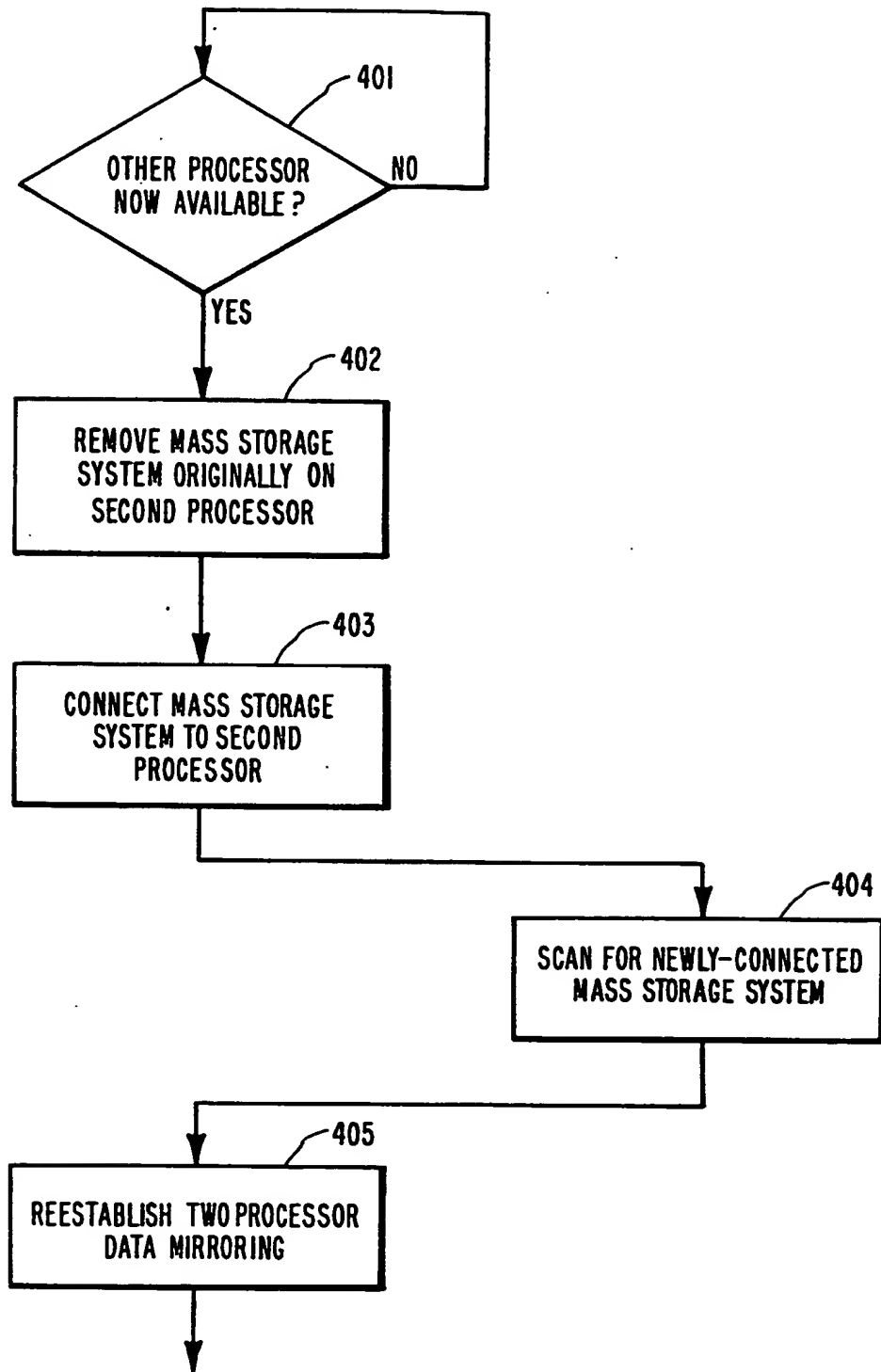


FIG. 4

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US94/07009

A. CLASSIFICATION OF SUBJECT MATTER

IPC(5) : G06F 11/34

US CL : 395/575

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 395/575; 371/8.1, 371/8.2, 371/9.1, 371/10.1, 371/10.2, 371/10.3, 371/11.1, 371/11.2, 371/11.3.

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

APS, US Automated Patent Searching System

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y, P	US, A 5,295,258 (Jewett et al.) 15 March 1994, col. 4 et seq., fig. 1, fig. 4.	1-24
Y	US, A, 5,157,663 (Major et al.) 20 October 1992, abstract et seq.	1-24



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:	* T	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
* A		document defining the general state of the art which is not considered to be part of particular relevance
* E		earlier document published on or after the international filing date
* L		document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
* O		document referring to an oral disclosure, use, exhibition or other means
* P		document published prior to the international filing date but later than the priority date claimed
	* X	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
	* Y	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
	* A	document member of the same patent family

Date of the actual completion of the international search

17 August 1994

Date of mailing of the international search report

SEP 02 1994

Name and mailing address of the ISA/US
Commissioner of Patents and Trademarks
Box PCT
Washington, D.C. 20231

Facsimile No. (703) 305-9564

Authorized officer

PHILLIP FRANCIS VALES

Telephone No. (703) 305-9683

**This Page is Inserted by IFW Indexing and Scanning
Operations and is not part of the Official Record**

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☐ BLACK BORDERS
- ☐ IMAGE CUT OFF AT TOP, BOTTOM OR SIDES
- ☐ FADED TEXT OR DRAWING
- ☐ BLURRED OR ILLEGIBLE TEXT OR DRAWING
- ☐ SKEWED/SLANTED IMAGES
- ☐ COLOR OR BLACK AND WHITE PHOTOGRAPHS
- ☐ GRAY SCALE DOCUMENTS
- ☐ LINES OR MARKS ON ORIGINAL DOCUMENT
- ☒ REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY
- ☐ OTHER: _____

IMAGES ARE BEST AVAILABLE COPY.

As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.